Animal 17 (2023) 101023

Contents lists available at ScienceDirect

Animal
The international journal of animal biosciences

# Measuring dairy cow welfare with real-time sensor-based data and farm records: a concept study

A.H. Stygar [a,*], L. Frondelius [b], G.V. Berteselli [d], Y. Gómez [c], E. Canali [d], J.K. Niemi [a], P. Llonch [c], M. Pastell [b]

[a] Bioeconomy and Environment, Natural Resources Institute Finland (Luke), Latokartanonkaari 9, 00790 Helsinki, Finland
[b] Production Systems, Natural Resources Institute Finland (Luke), Latokartanonkaari 9, 00790 Helsinki, Finland
[c] Department of Animal and Food Science, Universitat Autònoma de Barcelona, Campus UAB, 08193 Cerdanyola del Vallès, Barcelona, Spain
[d] Department of Veterinary Medicine and Animal Sciences, Università degli Studi di Milano, Via dell'Università 6, 26900 Lodi, Italy

ARTICLE INFO

*Article history:*
Received 1 June 2023
Revised 16 October 2023
Accepted 17 October 2023
Available online 27 October 2023

*Keywords:*
Accelerometer
Machine-learning
Monitoring
Precision livestock farming (PLF)
Welfare label

ABSTRACT

Welfare assessment of dairy cows by in-person farm visits provides only a snapshot of welfare and is time-consuming and costly. Possible solutions to reduce the need for in-person assessments would be to exploit sensor data and other routinely collected on-farm records. The aim of this study was to develop an algorithm to classify dairy cow welfare based on sensors (accelerometer and/or milk meter) and farm records (e.g. days in milk, lactation number). In total, 318 cows from six commercial farms located in Finland, Italy and Spain (two farms each) were enrolled for a pilot study lasting 135 days. During this time, cows were routinely scored using 14 animal-based measures of good feeding, health and housing based on the Welfare Quality® (**WQ®**) protocol. WQ® measures were evaluated daily or approximately every 45 days, using disease treatments from farm records and on-farm visits, respectively. WQ® measures were supplemented with daily temperature-humidity index to account for heat stress. The severity and duration of each welfare measure were evaluated, and the final welfare index was obtained by summing up the values for each cow on each pilot study day, and stratifying the result into three classes: good, moderate and poor welfare. For model building, a machine-learning (**ML**) algorithm based on gradient-boosted trees (XGBoost) was applied. Two model versions were tested: (1) a global model tested on *unseen* herd, and (2) a herd-specific model tested on *unseen* part of the data from the same herd. The version (1) served as an example on the model performance on a herd not previsited by the evaluator, while version (2) resembled a custom-made solution requiring in-person welfare evaluation for model training. Our results indicated that the global model had a low performance with average sensitivity and specificity of 0.44 and 0.68, respectively. For the herd-specific version, the model performance was higher reaching an average of 0.64 sensitivity and 0.80 specificity. The highest classification performance was obtained for cows in poor welfare, followed by cows in good and moderate welfare (balanced accuracy of 0.77, 0.71 and 0.68, respectively). Since the global model had low classification accuracy, the use of the developed model as a stand-alone system based solely on sensor data is infeasible, and a combination of in-person and sensor-based welfare evaluation would be preferable for a reliable welfare assessment. ML-based solutions, even with fair discriminative abilities, have the potential to enhance dairy welfare monitoring.



## Implications

This study provides insights into the potential application of machine-learning algorithms in dairy cow welfare assessment. We tested whether sensors and machine-learning algorithms can replace humans in classifying dairy cows into good, moderate and poor welfare classes. Our results show that humans cannot yet be substituted by machine learning. However, artificial intelligence can complement human evaluation. The welfare evaluation systems that integrate the human and machine-learning evaluation could improve the welfare monitoring of dairy cows, by providing real-time welfare evaluation. Solutions for a continuous welfare monitoring support efforts to achieve more socially acceptable dairy production.

* Corresponding author.
*E-mail address:* anna.stygar@luke.fi (A.H. Stygar).

https://doi.org/10.1016/j.animal.2023.101023
1751-7311/

## Introduction

Animal welfare is a complex concept (OIE, 2019), and verifying the welfare status of animals requires intricate monitoring tools. A recent market review of dairy welfare quality schemes revealed that current protocols are comprehensive in welfare assessment by covering various aspects of housing, feeding, health and behaviour of dairy cows (Stygar et al., 2022), but at the same time, welfare schemes are lagging in terms of the number of measures evaluated based on the animals and the utilisation of data routinely generated on farms.

Sensor technologies are commonly used on dairy farms in many European Union countries, with approximately 40–70% of farms using at least one technology (e.g. Lora et al., 2020; Utriainen et al., 2019). Also, the global market of dairy sensor technologies consists of a variety of products, with around 130 commercial sensors with potential application for animal-based welfare assessment (Stygar et al., 2021). Despite such a wide distribution of sensors, so far only one welfare quality scheme allows the direct application of sensor technologies for providing information, and even in this scheme only on one welfare measure, i.e. grazing time (Stygar et al., 2022). The main reason for the underutilisation of sensors for obtaining dairy cattle welfare information is likely the lack of algorithms aggregating information from several sensors and welfare domains to provide an individual welfare score.

Recent studies demonstrated that sensor data in combination with regression analyses or machine-learning (**ML**) algorithms can be used to estimate the complex traits of animals, such as the resilience rank of cows (Adriaens et al., 2020) or health status (Gertz et al., 2020). At the same time, works on algorithms using sensor data for providing animal welfare information for consumers and producers have been initiated (Llonch et al., 2021). Despite certain weaknesses of sensors, e.g. focus on measurable indicators rather than meaningful ones (Tuyttens et al., 2022), the application of precision livestock farming (**PLF**) solutions is an opportunity for continuous monitoring of animal welfare status, which can bring benefits for the whole dairy value chain.

Sensors provide the individual animal information on animal activity, behaviour and productivity, and therefore have the potential for animal-based welfare assessment (Stygar et al., 2021). However, to develop and validate a welfare classification algorithm using sensor data, a gold standard for animal welfare is needed. This gold standard can be based on existing protocols, such as the Welfare Quality protocol (**WQ®**) for dairy cows (Welfare Quality®, 2009). The original aggregation system of WQ® approach permits the integration of various measures (animal, management and resource-based) into an overall welfare classification on the herd level. However, for animal-based sensor application, a higher level of label granularity (increase from herd to individual level) is needed.

The main objective of this study was to develop and test the performance of a machine-learning algorithm based on gradient tree-boosting approach in classifying individual dairy cattle welfare status from sensor data and farm records. In this study, we sought a welfare assessment on individual animal level on a daily basis, therefore, animal-based measures from WQ® protocol were assigned with weight and duration, and summed up to obtain a daily animal welfare index. The proposed welfare index is an adaptation of WQ®.

## Material and methods

### *Welfare glossary*

The following terms and definitions were used in this study:

*Welfare measure* – measure taken on an animal that is used to assess a welfare indicator.
*Welfare index* – sum of all welfare measures taken on an animal, on a given day (ranging from 0 to 23).
*Welfare class* - synthesis of welfare index, allocating animal to a welfare category (good if welfare index < 2, moderate if welfare index = 2 or 3 and poor if welfare index > 3).
*Herd welfare rank* – rank of a herd within a grading system based on the average welfare class of all assessed animals in testing set.

### *Herds and data sources*

The study was conducted between 3 February 2021 and 17 June 2021 in six herds located in Finland, Spain and Italy. Altogether, 318 dairy cows were included in the study. The cows were selected to represent various lactation numbers and days in milk. The general description of the herds is presented in Table 1. Dairy cows

**Table 1**
The general information of dairy cattle farms enrolled for the study.

| Herd | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Country | Finland | Finland | Italy | Italy | Spain | Spain |
| Type of farming | C | O | C | C | C | C |
| Breed | HF, NR | HF, NR | HF | HF | HF | HF |
| Average number of dairy cows | 130 | 365 | 320 | 75 | 220 | 125 |
| Average number of heifers | 76 | 130 | 120 | 65 | 150 | 90 |
| Average length of productive life (months)[1] | 55 | 49 | 51 | 53 | 58 | 68 |
| Type of bedding (for milking cows)[2] | Wood shavings and mattress | Recycled manure solids and mattress | Straw and mattress | Straw and mattress | Compost[3] | Compost[3] and rice husk |
| Installed milking systems[4] | Parlour | AMS | Parlour | AMS | Parlour | Parlour |
| Access to pasture or outdoor area | No | Yes | Outdoor dry cows (1 month) | Outdoor dry cows (2 months) | No | No |
| Mortality rate (%) of dairy cows[5] | 4 | 4 | 2 | 5 | 2 | 2 |

Abbreviations: C = Conventional, O = organic, HF = Holstein Friesian, NR = Nordic Red, AMS = automatic milking system.
[1] Calculated as an average number of months from birth to slaughter for all milking cows in study farms during 2021.
[2] All herds were kept in free stall barns.
[3] Pile-matured dry manure from cows and heifers.
[4] Herd 1 used parlour system; however, data on individual milking records were not available.
[5] Percent of involuntary culling.

were kept in free stalls. During the study period, data with relevance to animal welfare were collected. The sensor data (accelerometers and milk recording devices) as well as farm records (e.g. lactation stage and parity of the cows) were used as predictor variables. The target variable (welfare class) was constructed using animal-based evaluation performed according to WQ® protocol, veterinary treatments and meteorological data (temperature, humidity).

### *Sensor data and farm records*

Neck-mounted accelerometers (Ida, Connecterra, the Netherlands) were deployed in all six herds. Time data obtained from accelerometers on lying, standing, walking, rumination, eating and other behaviours were available in 24-hour intervals. During the 24-hour time budget, individual cow data on lying, standing, walking, as well as rumination, eating and other behaviour sum up to 24. Data on other behaviour should be interpreted as activity not related to rumination or eating, e.g. drinking or inactivity.

Problems with missing data were recognised. This was mainly due to lost connection between the sensor and the receiver. Therefore, 6 days with accelerometers being switched off were removed. Additionally, observations for cows with daily lying time >1 000 minutes and <333 minutes were considered as outliers and excluded from the data set. In total, 422 observations for 149 cows were removed. The limits for the outliers were chosen based on the distribution of daily lying time presented in (Tucker et al., 2021).

Data on milk yield were collected from automatic milking systems (AMSs) installed in herd 2 (Finland; Lely Industries N.V., Maassluis, the Netherlands) and herd 4 (Italy; DeLaval Tumba, Sweden). In herd 3 (Italy), a parlour system was implemented using milk meters connected to Afimilk management system (Afimilk, Kibbutz Afikim, Israel). In herds 5 and 6 (Spain), the milk yield was recorded with GEA milking parlour (GEA Farm Technologies GmbH, Bönen, Germany). Herd 1 (Finland) did not have daily individual milk production data available. However, milk data were estimated using monthly milk recording system and historical data (method described in Supplementary Material S1).

Additionally, information on animals enrolled in the study, such as lactation number and days in milk (**DIM**), were collected from farm records. An overview of sensor features and farm record data is presented in Table 2. The mean and distribution of behavioural indicators of sensor data are shown in Fig. 1a and b.

### *Welfare assessment*

#### *On-farm welfare evaluation*

Trained welfare assessors visited the farms three times during the study period, on approximately days 1, 45 and 90. In each country, the evaluations were performed by the same assessor (authors L.F. herds 1–2, G.B. herds 3–4, Y.G. herds 5–6). On these visits, on-farm animal welfare data were collected according to the WQ® (Welfare Quality®, 2009) guidelines. Only animal-based measures that could be assessed at individual cow level were included. These measures covered three out of the four WQ® principles; good feeding (1 out of 2 measures), good housing (3 out of 8 measures) and good health (9 out of 13 measures). The avoidance distance test from the principle of appropriate behaviour had to be excluded from the data as in both Finnish farms (1 and 2) the feed bunk design hindered testing because of a lack of sufficient space. Descriptive statistics of the welfare measures are presented in Supplementary Table S1.

#### *Veterinary treatment records*

Farm records on animal health were used to complement the data obtained from the on-farm welfare evaluation to account for any health issues that could develop between farm visits. Each farmer was asked to provide farm records of disease diagnosis and treatments for all animals enrolled in the study for the duration of the trial. The health issues included in the data were: clinical mastitis, dystocia, respiratory diseases, reproductive diseases and metabolic disorders. Most of these measures are evaluated also in WQ®, e.g. recorded treatments for mastitis or dystocia are assessed at group level, whereas symptoms of respiratory diseases (nasal discharge) or reproductive diseases (vulvar discharge) are evaluated during the on-farm visits. A summary of all recorded treatments in the herds is presented in Supplementary Table S2.

#### *Meteorological data*

To incorporate information on environmental conditions and possible heat stress, we collected meteorological data from the locations of all farms between February and June 2021 (Visual Crossing Corporation, 2021). The meteorological data comprised the daily maximum temperature (**T**) (°C) and relative humidity (**RH**) (%), and these values were used to calculate the daily temperature-humidity index (**THI**) using the following equation (NRC, 1971):

$$THI = (1.8 \times T + 32) - [(0.55 - 0.0055 \times RH) \times (1.8 \times T - 26)]$$

Based on previous studies (Polsky and von Keyserlingk, 2017), we assumed that mild heat stress was present with indices between 71 and 79, and moderate to severe heat stress was present with indices >79.

#### *Welfare index and welfare class*

The aim of the welfare index, similarly as in the Welfare-Adjusted Life Year index (Teng et al., 2018), was to quantify the degree of impaired welfare on individual animal level. A daily animal welfare index was created by combining the on-farm animal welfare measures, farm records of diagnosed diseases and

**Table 2**
Overview of sensor features and farm records collected from six dairy cattle herds.

| Herd | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Number of cows included in the study | 52 | 36 | 56 | 51 | 60 | 63 |
| Average days in milk | 179 | 87 | 141 | 218 | 135 | 146 |
| Average lactation number | 2.2 | 1.5 | 1.8 | 2.2 | 2.1 | 2.2 |
| Number of daily sensors observations | 3 677 | 2 805 | 4 812 | 4 438 | 6 050 | 5 920 |
| Daily average milk production (kg) ± SD | 30.8 ± 6.7 | 33.02 ± 6.1 | 37.0 ± 7.2 | 24.4 ± 11.3 | 40.9 ± 5.5 | 39.8 ± 8.9 |
| Lying[1] (h) ± SD | 9.9 ± 1.1 | 10.4 ± 1.1 | 10.1 ± 1.3 | 10.7 ± 1.3 | 9.6 ± 1.3 | 10.8 ± 1.3 |
| Standing[1] (h) ± SD | 10.6 ± 1.2 | 10.8 ± 1.1 | 11.3 ± 1.4 | 10.2 ± 1.3 | 10.7 ± 1.4 | 10.4 ± 1.2 |
| Walking[1] (h) ± SD | 3.5 ± 0.9 | 2.9 ± 0.8 | 2.5 ± 0.8 | 3.1 ± 0.9 | 3.7 ± 0.8 | 2.7 ± 0.7 |
| Rumination[2] (h) ± SD | 4.3 ± 1.3 | 4.9 ± 1.5 | 4.9 ± 1.3 | 3.6 ± 1.0 | 3.7 ± 1.3 | 3.5 ± 1.2 |
| Eating[2] (h) ± SD | 6.8 ± 1.1 | 7.7 ± 1.0 | 8.0 ± 0.9 | 6.8 ± 1.2 | 7.8 ± 1.1 | 8.2 ± 1.2 |
| Other behaviour[2] (h) ± SD | 13.0 ± 1.8 | 11.5 ± 2.1 | 11.2 ± 1.8 | 13.6 ± 2.0 | 12.4 ± 1.8 | 12.2 ± 2.0 |

[1] Lying, standing, walking time daily observations for individual cows sum up to 24 hours.
[2] Rumination, eating, other behaviour time daily observations for individual cows sum up to 24 hours.

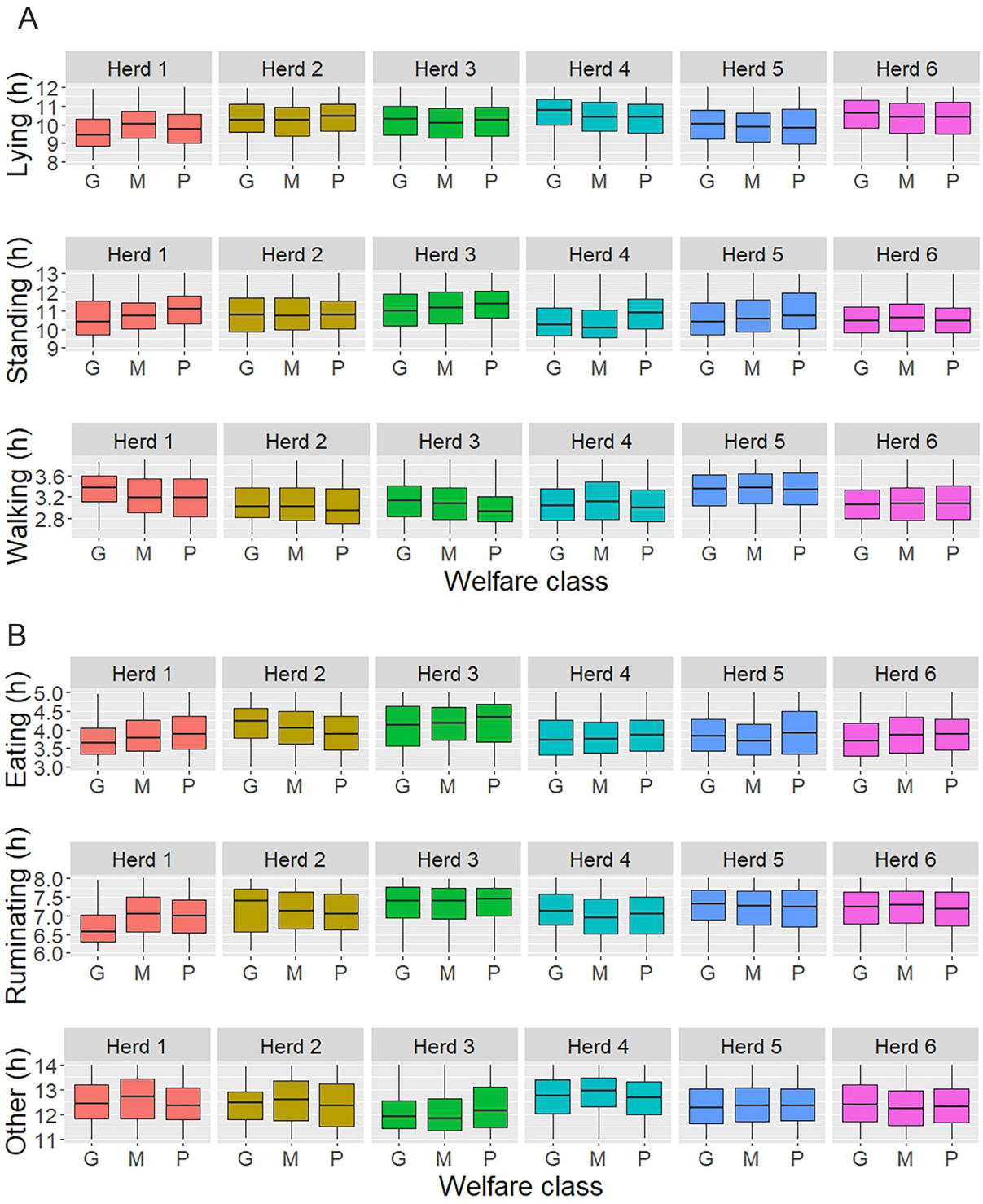


**Fig. 1.** Boxplot of (A) lying, standing, walking, (B) eating, ruminating and other behaviour time (in hours) for different dairy cattle herds (1–6) and various welfare classes. Abbreviations: G-good, M-moderate and P-poor.

meteorological data. The severity (0 = no welfare problem, 1 = mild to moderate welfare problem, 2 = severe welfare problem) and the duration (1–45 days) of the different welfare issues were estimated by the expert opinions of the authors and based on the relevant scientific literature. For those welfare measures that used information from both the WQ® and the farm records (e.g. respiratory diseases), the on-farm measure (e.g. nasal discharge) was allocated with severity rate 1 and the veterinary treatment (e.g. antibiotic treatment for pneumonia) was allocated with severity rate 2. The final animal welfare index for each cow was obtained by summing up the severity scores on daily basis. A cow could receive a score from 0 to a maximum of 23. The example of an index for a cow that developed severe lameness and deteriorated in terms of cleanliness and integument alterations is presented on Fig. 2. Based on the index, each cow was classified in a good (0–1), moderate (2–3), or poor (>3) welfare class. Descriptions of the welfare measures,

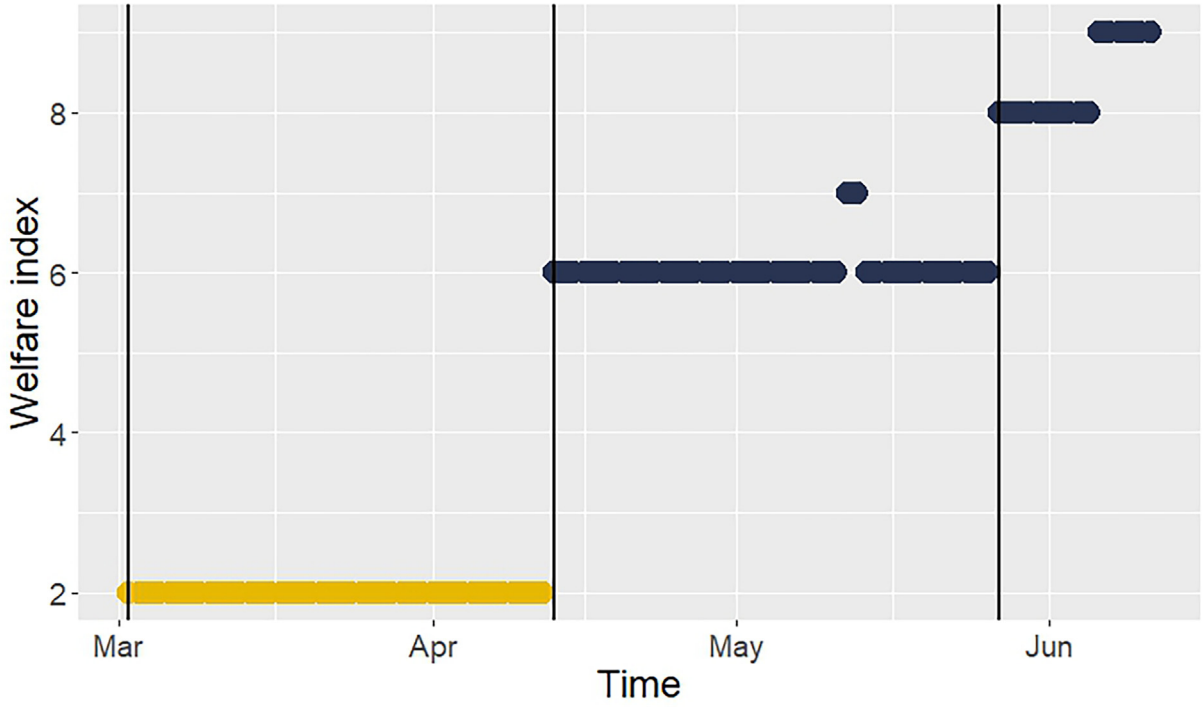


**Fig. 2.** The welfare index obtained from an example cow which had problems with leg cleanliness and moderate integument alterations (first Welfare Quality® assessment), but later developed lameness and severe integument alterations (second Welfare Quality® assessment), which further deteriorated to problems with leg, udder and flank cleanliness, severe integument alterations and lameness (third Welfare Quality® assessment). The colours denote the welfare class, with moderate (2–3) marked in yellow and poor (>3) marked in blue. The vertical lines indicate Welfare Quality® scoring days.

their severity and duration are presented in Table 3. The plot of welfare index per farm is shown in Fig. 3. The total dataset contained 8 166 good, 11 478 moderate and 8 108 poor welfare classes. To assess how much welfare class changed after WQ® evaluation, we run paired sample t-test statistics on subsequent observations (the day proceeding WQ® assessment and the day with WQ® assessment).

*Model building*

Two versions of the model were tested in this study: (1) a global model tested on an *unseen* herd and (2) a herd-specific model tested on an *unseen* part of the data from the same herd. The general-purpose model served as an example on how the model performs on a herd not previsited by the evaluator, while the herd-specific model represents a scenario where information on animal welfare gathered by human observer for this given herd is available in the model training phase. The performance of the models was verified using a cross-validation procedure. For a general-purpose model, 6-fold cross-validation resampling based on herd number was applied (data from 5 herds used in training set, data from 1 herd used in testing set). In the herd-specific model, 3-fold cross-validation based on time after WQ assessment was implemented and model data from each herd were divided into 15-day intervals. The total time was 45 days, of which 30 days were used for training and 15 for testing.

*Sensor features processing*

Data used for model building and testing, after outliers' removal, were preprocessed with smoothing and normalisation procedures. The smoothing procedure based on moving average was applied to remove noise (e.g., measurement error) from the time series sensor data (milk meter and accelerometer recordings). The normalisation was implemented to ensure that data between farms were comparable.

The smoothing procedure assumed selection of window widths (in this study, windows of 5, 7 and 10 days were tested) and obtaining an average and SD estimate. Additional feature processing concerned calculating a linear regression for different time windows (5, 7 and 10 days) and storing the slope as well as obtaining the difference between the average in various time windows (between features means from 7- and 5-day windows as well as 10- and 5-day windows). Therefore, for example, if a decreasing pattern in a cow's lying time was observed over a period of 5, 7 or 10 days, the slope and difference feature will be negative, and if the opposite was observed, the slope and difference feature will be positive.

Two different normalisation procedures were tested, namely daily rank and Z-score. Both procedures were applied to individual measurements within each herd and resulted in different distribution of variables. Rank was based on rank transformation of the sensor features within herd for each day resulting in values between 0 and 1. The Z-score normalisation was performed using the following equation:

$$Z = \frac{x - \bar{X}}{\sigma}$$

where $\bar{X}$ is the daily mean and $\sigma$ is the daily SD of a given sensor feature in each herd.

*Machine-learning algorithm*

The XGBoost classification algorithm based on the gradient tree-boosting approach (Chen et al., 2022) was applied. The objective of the model was multi-class classification ('multi:softprob') which allowed the categorisation of the test data into the multiple labels of good, moderate and poor welfare. The learning task was

**Table 3**
Measures from on-farm animal welfare assessment, veterinary treatment data and meteorological data used to obtain daily individual-level dairy cow welfare index, grouped according to Welfare Quality® principles.

| Welfare principle | Measure | Severity | Duration (days)[1] | Reference |
|---|---|---|---|---|
| Good feeding | Body condition score (normal, very fat, very lean)[2] | 0/1/2 | 45 | Roche et al., 2009 |
| Good housing | Cleanliness, udder (clean/dirty)[2] | 0/1 | 45 | Sant'Anna and Paranhos da Costa, 2011 |
| | Cleanliness, legs (clean/dirty)[2] | 0/1 | 45 | |
| | Cleanliness, flank (clean/dirty)[2] | 0/1 | 45 | |
| | Thermal comfort[3] (no heat stress/heat stress/ severe heat stress) | 0/1/2 | 1 | Polsky and von Keyserlingk, 2017 |
| Good health | Integument alterations (not present/hairless patch/lesion and/or swelling)[2] | 0/1/2 | 45 | Frondelius et al., 2020; Vokey et al., 2001 |
| | Locomotion score (not lame, lame, severely lame)[2] | 0/1/2 | 45 | Groenevelt et al., 2014; Hoblet and Weiss, 2001 |
| | Mastitis (not present/treatment for clinical mastitis)[4] | 0/2 | 21 | Fogsgaard et al., 2015 |
| | Respiratory disease (not present/ nasal discharge based on Welfare Quality®/treatment for respiratory disease)[5] | 0/1/2 | 7 | Smith, 2014 |
| | Ocular discharge (not present/present)[2] | 0/1 | 7 | Smith, 2014 |
| | Reproductive disease (not present/ vulvar discharge based on Welfare Quality®/treatment for inflammatory reproduction disease)[5] | 0/1/2 | 7 | Neave et al., 2018 |
| | Dystocia[4] (not present/ treatment for dystocia) | 0/2 | 7 | Smith, 2014 |
| | Diarrhea[2] (not present/present) | 0/1 | 7 | Smith, 2014 |
| | Metabolic disorder (not present/treatment for metabolic disorder)[4] | 0/2 | 7 | Stangaferro et al., 2016 |

[1] Depending on the measure, the duration will start from the day of Welfare Quality® assessment or the day of treatment registration in farm records.
[2] Based on Welfare Quality® assessment performed on cows enrolled in the study.
[3] Based on temperature humidity index, heat stress was present between indices 71–79, severe heat stress was present with indices >79.
[4] Based on farm records (treatment records or notes with diagnosis).
[5] Based on both Welfare Quality® evaluation and farm records (treatments records were checked, if cow received antimicrobial treatment due to respiratory disease or reproduction disease vulvar or nasal discharge was changed to 2).

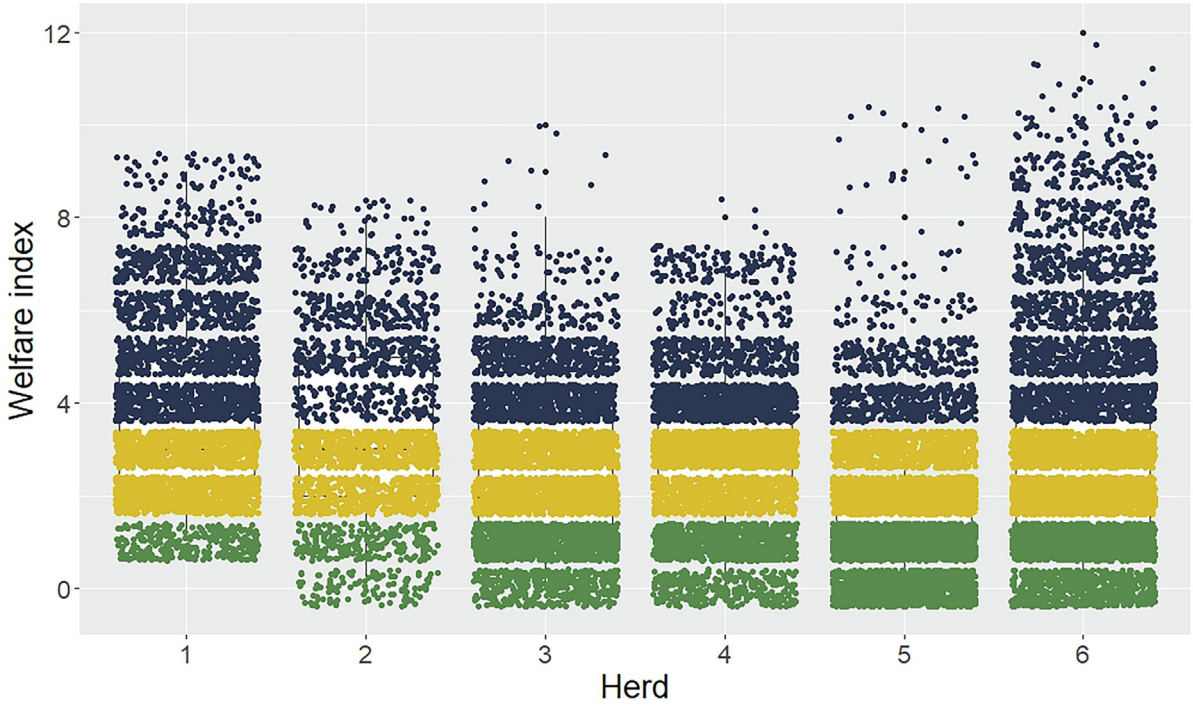


**Fig. 3.** The welfare index obtained in six herds for all enrolled cows during the study period. Colours indicate various welfare classes: green colour for good (index < 2), yellow for moderate (index = 2 or 3) and blue for poor (index > 3) welfare.

set as a non-linear tree algorithm ("gbtree") with suitable booster parameters. The model hyperparameters were tuned using a "random search" approach on training set. To implement the random search, firstly, we determined the range of values and increments of the relevant parameters (Supplementary Table S3). Secondly, a random set of parameters was sampled, used for model building in a ten-fold cross-validation based on cows' identification number and evaluated based on classification performance. The procedure was repeated for n = 100 times. The parameter set with the highest predictive performance (accuracy) was selected for final model testing.

To address the problem of unbalanced data set, namely a low proportion of animals with good or poor welfare, compared to moderate welfare, as well as to increase importance of observa-

tions closer to the assessment dates, a weight vector was defined and used as instance weights in models cost function (Chen et al., 2022). The weight vector was calculated using the following equation:

$$weight_{i,j} = (1 - 1/45 DayNr) \times \left(\frac{n_j}{n}\right)$$

where $weight_{i,j}$ is assigned to observation from cow $i$ with welfare class $j$, given the information on total number of animals ($n$), number of animals in particular welfare class $n_j$, as well as the number of days after the on-farm assessment visit ($DayNr$). Constant numbers from the equation were used to calculate decreasing importance between day 0 and day 45 after the welfare assessment. For example, observations made on days 0, 10 and 45 after the WQ® assessment were given weights of 1.00, 0.78 and 0.01, respectively. Processed sensor features were supplemented with data on lactation stage (days in milk) and lactation number.

*Model validation*

The multiclass XGBoost algorithm returns the probability values of a cow being in a specific welfare class (good, moderate or poor). The classification performance of the algorithm was assessed using several measures: sensitivity, specificity, multiclass area under the receiver operating characteristic curve (**AUC**), and balanced accuracy (**BA**). The BA was calculated simply as the arithmetic mean of sensitivity and specificity in predicting good, moderate and poor welfare.

The classification results obtained in each tested scenario concerning model type, smoothing and normalisation procedure were presented as mean and SD. Additionally, the mean value of the fractional contribution of each feature to the model based on the total gain was calculated. Finally, the predictions from the model were used to estimate the herd welfare rank.

Data management, plotting, model building and testing were done using the R (R Core Team, 2017) language version 4.2.1 extended with XGBoost (Chen et al., 2022) and pRoc (Robin et al., 2011) packages. All calculations were done using standard laptop 64bit Windows 10 computer equipped with Intel® Core™ i5-1235U processor 4.4GHZ and 16 GB RAM.

## Results

Table 4 presents the actual average welfare obtained in all herds. The worst welfare was found in herd 1 (welfare class was on average 2.4 ± 0.6), while the best welfare was observed in herd 5 (on average 1.6 ± 0.7). The highest variation in the welfare class was detected in the herd 6 (SD = 0.8). Paired sample t-test statistics for welfare classes on day proceeding WQ® assessment and day with WQ® assessment are presented in Supplementary Table S4. The results indicate significant differences between two evaluation days in six out of 12 analysed cases.

*Model performance*

Averaged results obtained from two model versions run under different smoothing and normalisation strategies are presented in Table 5. The results show that the classification ability for the general-purpose model is close to a random classifier, with the average sensitivity and specificity of 0.41 and 0.68, respectively, and an AUC value of 0.52. The performance of the herd-specific models varied slightly depending on the data preprocessing strategy, but for selected normalisation and smoothening strategies model reached acceptable discrimination ability (AUC $\geq$ 0.70, (Hosmer et al., 2013)).

The highest classification performance (AUC = 0.70) was achieved using a 5-day time window for the smoothing strategy and Z-score transformation as the normalisation strategy. The best classification performance was obtained in the herd 1 (AUC = 0.77), while the lowest performance evaluation was reached in herd 6 (AUC = 0.63). Based on the BA values calculated for each welfare class, the highest classification performance was obtained for cows in poor welfare, followed up by cows in good welfare (BA of 0.77 and 0.71, respectively). Cows in moderate welfare classes were often misclassified by the model (BA of 0.67).

**Table 4**
Welfare evaluation (actual, predicted and herd ranking) for all dairy cattle herds in the study. The predicted values of herd welfare were obtained using herd-specific welfare classification model Z-score 5.

| Herd number | Actual welfare[1] | Predicted welfare[2] | Herd ranking (actual-predicted)[3] |
|---|---|---|---|
| 1 | 2.39 ± 0.63 | 2.27 ± 0.55 | 6–6 |
| 2 | 2.18 ± 0.68 | 2.11 ± 0.50 | 5–5 |
| 3 | 1.94 ± 0.76 | 1.79 ± 0.56 | 2–2 |
| 4 | 2.10 ± 0.75 | 2.09 ± 0.70 | 4–4 |
| 5 | 1.60 ± 0.68 | 1.36 ± 0.62 | 1–1 |
| 6 | 2.04 ± 0.78 | 1.98 ± 0.50 | 3–3 |

[1] An average ± SD welfare class in testing data set.
[2] An average ± SD predicted welfare class.
[3] Herds were ranked from the best (1) to the worst welfare (6).

**Table 5**
Results of model performance dairy cow welfare classification for different normalisation procedures and testing options (mean ± SD).

| Tested model versions | Smoothening and normalisation strategy[1] | Sensitivity | Specificity | AUC | Balanced accuracy for welfare class | | |
|---|---|---|---|---|---|---|---|
| | | | | | Good welfare | Moderate welfare | Poor welfare |
| The global model (the 6-fold cross validation, train on five herds, test on remaining) | None 0 | 0.42 ± 0.22 | 0.69 ± 0.14 | 0.53 ± 0.04 | 0.51 ± 0.07 | 0.54 ± 0.09 | 0.61 ± 0.13 |
| | Rank 5 | 0.39 ± 0.22 | 0.68 ± 0.13 | 0.53 ± 0.04 | 0.48 ± 0.06 | 0.51 ± 0.07 | 0.61 ± 0.13 |
| | Rank 7 | 0.43 ± 0.22 | 0.68 ± 0.13 | 0.53 ± 0.03 | 0.52 ± 0.05 | 0.52 ± 0.06 | 0.62 ± 0.13 |
| | Rank 10 | 0.41 ± 0.23 | 0.69 ± 0.13 | 0.53 ± 0.03 | 0.52 ± 0.04 | 0.52 ± 0.03 | 0.60 ± 0.12 |
| | Z-score 5 | 0.43 ± 0.23 | 0.70 ± 0.12 | 0.52 ± 0.02 | 0.53 ± 0.07 | 0.54 ± 0.05 | 0.62 ± 0.13 |
| | Z-score 7 | 0.42 ± 0.23 | 0.68 ± 0.13 | 0.52 ± 0.04 | 0.52 ± 0.07 | 0.52 ± 0.05 | 0.61 ± 0.13 |
| | Z-score10 | 0.41 ± 0.24 | 0.68 ± 0.12 | 0.51 ± 0.03 | 0.53 ± 0.11 | 0.52 ± 0.07 | 0.58 ± 0.15 |
| The herd-specific welfare classification model (the 3-fold cross-validation) | None 0 | 0.61 ± 0.18 | 0.78 ± 0.10 | 0.66 ± 0.07 | 0.66 ± 0.11 | 0.65 ± 0.08 | 0.78 ± 0.06 |
| | Rank 5 | 0.62 ± 0.16 | 0.79 ± 0.09 | 0.68 ± 0.06 | 0.67 ± 0.10 | 0.66 ± 0.07 | 0.76 ± 0.06 |
| | Rank 7 | 0.63 ± 0.16 | 0.79 ± 0.10 | 0.68 ± 0.07 | 0.70 ± 0.10 | 0.67 ± 0.08 | 0.76 ± 0.07 |
| | Rank 10 | 0.62 ± 0.16 | 0.79 ± 0.10 | 0.69 ± 0.08 | 0.69 ± 0.11 | 0.66 ± 0.09 | 0.75 ± 0.07 |
| | Z-score 5 | 0.64 ± 0.17 | 0.80 ± 0.10 | 0.70 ± 0.09 | 0.71 ± 0.12 | 0.68 ± 0.09 | 0.77 ± 0.07 |
| | Z-score 7 | 0.62 ± 0.17 | 0.79 ± 0.10 | 0.68 ± 0.07 | 0.69 ± 0.10 | 0.66 ± 0.08 | 0.76 ± 0.06 |
| | Z-score 10 | 0.63 ± 0.16 | 0.79 ± 0.10 | 0.69 ± 0.06 | 0.71 ± 0.10 | 0.67 ± 0.08 | 0.75 ± 0.06 |

Abbreviations: AUC = area under the operating characteristic curve.
[1] Number refers to the length of the time window used for calculating moving average.

The herd welfare rank based on actual and predicted values (using herd-specific welfare classification model Z-score 5) is presented in Table 4. Even though, on average, the model predicted herds better welfare than was observed, all herds were ranked correctly with herd 1 having the worst welfare and herd 5 having the best welfare.

The results of the XGBoost hyperparameter optimisation for the three best models are shown in Supplementary Table S5. The Eta and subsample means were almost identical for all three models. The largest differences between the tested versions were observed for the parameter *nrounds*, which indicates the number of decision trees in the final model and *Gamma* which specifies the minimum loss reduction required to make a split.

The average computation time for the training and testing phase in herd-specific model for single fold was 2 minutes 28 seconds. However, this step did not include the time needed to preprocess the data (data loading, normalisation and smoothing).

### *The most important predictor variables*

The most important predictor variables (the first five features with the highest relevant contribution) for all tested models are presented in Supplementary Table S6. For the general-purpose model, the most frequently selected variables were lactation number, daily milk yield, DIM, walking and the slope of linear regression for milk yield. In the herd-specific version daily milk yield, DIM, lactation number, walking and other behaviour time had the highest relative importance in the model.

## Discussion

In this study, two ML-based algorithms were constructed and tested: (1) a global model tested on an *unseen* herd, and (2) a herd-specific model tested on an *unseen* part of the data from the same herd. Our results indicated that the global model had a low performance (AUC = 0.53 ± 0.03), whereas the herd-specific model version reached an acceptable discrimination ability (AUC = 0.70 ± 0.09). The best classification performance was obtained for cows in poor welfare, followed up by cows in good welfare. Despite the misclassification of individual animals (especially for animals with moderate welfare), predictions obtained from the model allowed accurate herd ranking. However, with regard to herd ranking, the results must be interpreted with caution due to the small herd sample size and relatively short observation period.

As an example of the difficulties in establishing a common model structure, Adriaens et al. (2020) reported the lack of a common trend for predicting dairy cattle resilience based on sensor data, while Naqvi et al. (2022) demonstrated reduced performance of mastitis detection models with increased variability in milk production. Cow behaviour is related to welfare. For example, lying time is higher in lame cows and lower in those with mastitis (Tucker et al., 2021). In our study, lying time increased in some herds as welfare deteriorated, but the opposite pattern was observed in other herds (Fig. 1a). This inconsistency may explain why behavioural features have relatively low importance in the model and why finding a common model across herds for welfare prediction remains a challenge.

Cows' welfare is usually evaluated using more than two classes (e.g. Welfare Quality®, 2009); therefore, the algorithm presented in this paper was defined as a multi-class classification problem. Whereas the limits of the welfare classes were chosen to obtain a balanced data set, the final number and limits of welfare classes should be selected taking into account possible trade-offs in classification outcomes, as demonstrated for example in simulation studies on lameness management (Edwardes et al., 2023). Our results showed that the model classification performance for cows in moderate welfare was low, which could be caused by no changes in behaviour or production for animals with mild welfare problems (e.g., cows with cleanliness issues) or due to measurement error originating from sensor data. Similar difficulties with misclassification of animals with moderate welfare problems have been noted in previous studies predicting lameness scores (e.g. Frondelius et al., 2022; Garcia et al., 2014). In order to enhance practical relevance, the model's accuracy should be further improved. For example, to increase the predictive power of behavioural data, further information on the reproduction status of animals (e.g. estrus detection) could be included as model variables. Accurate data labelling is essential for the performance of algorithms. The comparison between the welfare classes obtained on the day proceding the WQ® assessment and the day when WQ® assessment was performed showed significant differences between the group means of the welfare classes for select herds and assessment days. This result may indicate a certain degree of imprecision in the welfare label. The approach adopted in this study, concerning the frequency of farm visits required to obtain a welfare label, was a compromise between accuracy and cost of the assessment. To compensate for potentially imprecise labelling, we implemented a weighting strategy by increasing the contribution of the welfare labels closer to the assessment days to the loss function. In order to reduce the problem of inconsistent data annotations, future efforts should focus on combining human and automated data labelling approaches. For more accurate labelling, human evaluation could be supplemented by alarms generated by sensor systems other than those used for welfare prediction (e.g. sensors monitoring milk properties).

Recently, Gertz et al. (2020) demonstrated that the XGBoost classification approach can be successfully implemented to predict "sick" and "healthy" cows. When it comes to the herd-specific model version for welfare classification, the implementation of XGBoost yielded fair performance results. There might be several metrics to measure the feasibility of ML models for welfare assessment, but in the case of this study, perhaps the most obvious would be to compare the accuracy (for individual animal and herd ranking) and speed of human evaluation with the ML-based approach. In-person animal welfare assessments are not free from errors (e.g. Czycholl et al., 2018). Human performance in assessing animal welfare may vary and depend on, among others, the experience of the assessor (Katzenberger et al., 2020) or the complexity of the assessment (Schlageter-Tello et al., 2015). Therefore, further testing is needed to provide evidence on the performance of the model in comparison to human observers with different levels of experience. A major difference between human and ML-based welfare evaluation is in the time needed to perform the evaluation. While the collection and processing of the multiple measures in a single herd by a human observer can take several hours (Welfare Quality®, 2009), application of the ML algorithm, once the sensors are in place on the farm, could reduce this time to few minutes. After the necessary training period with on-farm in-person evaluation, welfare classification can be done in real-time, as soon as the data from the herd is available for the model inputs, and in a continuous manner. Therefore, even though the presented models reached only acceptable classification abilities, the tools may still be useful in practical conditions, for example by providing day-to-day welfare assessment of individual cows.

One of the main aims of this study was to design a flexible modelling framework for animal welfare classification. As there might be more than one welfare definition, e.g. due to legal requirements or consumer preferences (Stygar et al., 2022), we acknowledge that the model labelling method can change based on the user preferences. Therefore, in the case of this study, as important as obtaining results, was the testing process and the opportunity to learn

about the unexplored concept of model-based welfare assessment. This study helped us to pinpoint potential barriers and difficulties in the development of sensor-based welfare assessment.

At this stage of the development, equally important to the model performance statistics are design policies which could facilitate the successful implementation of the welfare monitoring platform in the commercial settings. We have identified four key elements which might impact the implementation of ML-based welfare assessment, namely (1) model robustness, (2) welfare definition, (3) missing predictors, and (4) stakeholder engagement in the designing phase. Each of these elements is discussed below.

The results of this study suggest that due to the substantial differences between herds, dairy cattle welfare classification on an unseen herd seems to be unfeasible and a training period with label data is necessary to improve model predictions. It is possible that by significantly increasing the number of farms and the duration of training data collection in each farm for training of the algorithm, better accuracy could be achieved. Anyhow, stand-alone systems based on sensor data might still be inaccurate and a combination of in-person and algorithm-based welfare evaluation will help to develop a more reliable, real-time dairy cattle welfare assessment. The direct consequence of low performance of the general-purpose model is an increase in the costs of assessment, as farms would need to be visited by evaluators to obtain initial information on welfare status. On the other hand, one might argue that the evaluation visits are unavoidable. The initial training period on herd-specific data might be necessary in case the algorithm becomes a part of a larger welfare evaluation platform using devices from different suppliers. In addition, the calibration periods with in-person evaluation might be necessary due to behavioural deviations over time, caused by substantial changes in farm management, such as adjustments in animal grouping. Finally, the human-ML approach can be a preferable solution for consumers who have expressed concerns about the implementation of various sensor technologies for welfare monitoring (Krampe et al., 2021).

In the current study, WQ® (Welfare Quality®, 2009), which is the most widely used and scrutinised animal welfare assessment protocol (Brscic et al., 2021; Tuyttens et al., 2021), was adopted. Nevertheless, when collecting data, difficulties with measuring the appropriate behaviour measures, which are part of the welfare definition according to WQ®, were encountered. Therefore, the nature of used methodology for data labelling (welfare definition) might bring some difficulties in constructing the general method for welfare algorithms applicable across European dairy farms. As previous studies have criticised measures defined inappropriate behaviour, for being insufficiently reliable for welfare assessment (e.g. Bokkers et al., 2012), future efforts should focus on providing simplified methods for measuring welfare.

When considering missing predictors, this study assumed the availability of sensor data and the approach was therefore targeted at farms using sensor technologies. In practice though, dairy companies deal with farms with varying degrees of digitalisation. Therefore, a strategy to handle missing sensor inputs needs to be developed. In our study, one farm did not record individual milk production. However, missing daily milk yields were extrapolated from monthly farm milk yield test records. As a consequence, the estimated milk yield data have a reduced variability and will not indicate temporal pattern change due to welfare issues. In the future, in case of missing sensor data, the potential of using alternative data sources, such as bulk tank registrations, should be explored. The wider availability of sensor features (from camera-based sensor technologies) might increase model precision but would also limit the proposed animal welfare assessment methodology to only highly automated farms.

The system should be designed to meet the needs of the end users. For farmers, model precision and ability to provide feedback, similarly as in WQ® (Roe et al., 2011), might be an important element determining their interest towards on-farm application of the system. Hence, it is essential for a welfare evaluation platform to gather end-user opinions on the desired functionality of the system. Consultations should be carried out with all interest groups, including farmers, consumers, NGOs and companies, to explore the potential of data-based solutions in providing animal welfare information.

## Conclusions

The main objective of this paper was to develop and test the performance of a machine-learning algorithm in classifying dairy cattle welfare status from sensor data and farm records. The final evaluation of the performance of the proof-of-concept model for welfare assessment will depend on the purpose of use by the end users. If the model is intended as the tool for farmers' decision support, accuracy should be improved to provide a reliable tool for effective animal welfare improvements. On the other hand, for identifying herds with specific welfare status for benchmarking or labelling purpose, a minimum performance may have been already achieved.

Machine-learning-based solutions, even with fair discriminative ability, have the potential to enhance the level of dairy welfare monitoring, which is currently a concern across European Union States.

## Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.animal.2023.101023.

## Ethics approval

Not applicable.

## Data and model availability statement

Data are not available due to the General Data Protection Regulation (GDPR). Model available on request.

## Declaration of Generative AI and AI-assisted technologies in the writing process

The authors did not use any artificial intelligence-assisted technologies in the writing process.

## Author ORCIDs

**A.H. Stygar:** https://orcid.org/0000-0003-3112-2847
**J.K. Niemi:** https://orcid.org/0000-0002-9545-3509
**P. Llonch:** https://orcid.org/0000-0001-5958-861X
**G.V. Berteselli:** https://orcid.org/0000-0002-2896-3971
**E. Canali:** https://orcid.org/0000-0002-4962-5955
**L. Frondelius:** https://orcid.org/0000-0002-5813-8948
**M. Pastell:** https://orcid.org/0000-0002-5810-4801

## Author contributions

AS, MP, JN, EC, and PL developed the study concept. LF, GB and YG collected data on farms (WQ assessment). MP and AS merged

data, defined and build model structure. AS run analyses and drafted the text of the paper. All authors contributed to the article and approved the submitted version.

## Declaration of interest

None.

## Acknowledgments

We would like to acknowledge Diego E Ruiz Di Genova from COVAP as well as six anonymous farmers who have provided us access to the data collected on farms as well as supported the welfare assessment data collection during the study period. The article was deposited as a preprint in a preprint repository (Stygar et al., 2023).

## Financial support statement

This study was conducted within the ClearFarm Project aiming to co-design, develop, and validate a software platform powered by PLF Technologies to provide animal welfare information. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 862919. PL received financial support from the Ramón y Cajal programme (RYC2020-029067-I) funded by the Spanish Ministry of Science and Innovation.

## References

Adriaens, I., Friggens, N.C., Ouweltjes, W., Scott, H., Aernouts, B., Statham, J., 2020. Productive life span and resilience rank can be predicted from on-farm first-parity sensor time series but not using a common equation across farms. Journal of Dairy Science 103, 7155–7171. https://doi.org/10.3168/jds.2019-17826.

Bokkers, E., de Vries, M., Antonissen, I., de Boer, I., 2012. Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. Animal Welfare 21, 307–318. https://doi.org/10.7120/09627286.21.3.307.

Brscic, M., Contiero, B., Magrin, L., Riuzzi, G., Gottardo, F., 2021. The use of the general animal-based measures codified terms in the scientific literature on farm animal welfare. Frontiers in Veterinary Science 8, 634498. https://doi.org/10.3389/fvets.2021.634498.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2022. xgboost: Extreme Gradient Boosting. Retrieved on 11 April 2022 from https://github.com/dmlc/xgboost.

Czycholl, I., Kniese, C., Casey-Schrader, L., Krieter, J., 2018. How reliable is the multi-criteria evaluation system of the Welfare Quality® protocol for growing pigs? Animal Welfare 27, 147–156. https://doi.org/10.7120/09627286.27.2.147.

Edwardes, F., Van Der Voort, M., Hogeveen, H., 2023. Quantifying the economic and animal welfare tradeoffs of classification models in precision livestock farming for sub-optimal mobility management (preprint). SSRN. https://doi.org/10.2139/ssrn.4511088, Published online by SSRN 15 July 2023.

Fogsgaard, K.K., Bennedsgaard, T.W., Herskin, M.S., 2015. Behavioral changes in freestall-housed dairy cows with naturally occurring clinical mastitis. Journal of Dairy Science 98, 1730–1738. https://doi.org/10.3168/jds.2014-8347.

Frondelius, L., Lindeberg, H., Pastell, M., 2020. Recycled manure solids as a bedding material: Udder health, cleanliness and integument alterations of dairy cows in mattress stalls. Agricultural and Food Science 29, 420–431. https://doi.org/10.23986/afsci.95603.

Frondelius, L., Lindeberg, H., Pastell, M., 2022. Lameness changes the behavior of dairy cows: daily rank order of lying and feeding behavior decreases with increasing number of lameness indicators present in cow locomotion. Journal of Veterinary Behavior 54, 1–11. https://doi.org/10.1016/j.jveb.2022.06.004.

Garcia, E., Klaas, I., Amigo, J.M., Bro, R., Enevoldsen, C., 2014. Lameness detection challenges in automated milking systems addressed with partial least squares discriminant analysis. Journal of Dairy Science 97, 7476–7486. https://doi.org/10.3168/jds.2014-7982.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., Krieter, J., 2020. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. Computers and Electronics in Agriculture 173, 105404. https://doi.org/10.1016/j.compag.2020.105404.

Groenevelt, M., Main, D.C.J., Tisdall, D., Knowles, T.G., Bell, N.J., 2014. Measuring the response to therapeutic foot trimming in dairy cows with fortnightly lameness scoring. The Veterinary Journal 201, 283–288. https://doi.org/10.1016/j.tvjl.2014.05.017.

Hoblet, K.H., Weiss, W., 2001. Metabolic Hoof Horn Disease Claw Horn Disruption. Veterinary Clinics of North America: Food Animal Practice 17, 111–127. https://doi.org/10.1016/S0749-0720(15)30057-8.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. Wiley series in probability and statistics,. Wiley, Hoboken, NJ, USA.

Katzenberger, K., Rauch, E., Erhard, M., Reese, S., Gauly, M., 2020. Inter-rater reliability of welfare outcome assessment by an expert and farmers of South Tyrolean dairy farming. Italian Journal of Animal Science 19, 1079–1090. https://doi.org/10.1080/1828051X.2020.1816509.

Krampe, C., Serratosa, J., Niemi, J.K., Ingenbleek, P.T.M., 2021. Consumer perceptions of precision livestock farming—A qualitative study in three European countries. Animals 11, 1221. https://doi.org/10.3390/ani11051221.

Llonch, P., Ingenbleek, P.T.M., Bokkers, E., Canali, E., Gosliga, S.P.V., Baxter, A., Manteca, X., 2021. The ClearFarm project: developing a platform to assess animal welfare using sensor technology in pigs and dairy cattle. In: Boyle, L., O'Driscoll, K. (Eds.), Proceedings of the 8th International Conference on the Assessment of Animal Welfare at Farm and Group level. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 170–170.

Lora, I., Gottardo, F., Contiero, B., Zidi, A., Magrin, L., Cassandro, M., Cozzi, G., 2020. A survey on sensor systems used in Italian dairy farms and comparison between performances of similar herds equipped or not equipped with sensors. Journal of Dairy Science 103, 10264–10272. https://doi.org/10.3168/jds.2019-17973.

Naqvi, S.A., King, M.T.M., DeVries, T.J., Barkema, H.W., Deardon, R., 2022. Data considerations for developing deep learning models for dairy applications: A simulation study on mastitis detection. Computers and Electronics in Agriculture 196, 106895. https://doi.org/10.1016/j.compag.2022.106895.

Neave, H.W., Lomb, J., Weary, D.M., LeBlanc, S.J., Huzzey, J.M., von Keyserlingk, M.A.G., 2018. Behavioral changes before metritis diagnosis in dairy cows. Journal of Dairy Science 101, 4388–4399. https://doi.org/10.3168/jds.2017-13078.

NRC, 1971. A Guide to Environmental Research on Animals. The National Academy of Sciences, Washington, DC, USA.

Polsky, L., von Keyserlingk, M.A.G., 2017. Invited review: Effects of heat stress on dairy cattle welfare. Journal of Dairy Science 100, 8645–8657. https://doi.org/10.3168/jds.2017-12651.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria. https://www.R-project.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.

Roche, J.R., Friggens, N.C., Kay, J.K., Fisher, M.W., Stafford, K.J., Berry, D.P., 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. Journal of Dairy Science 92, 5769–5801. https://doi.org/10.3168/jds.2009-2431.

Roe, E., Buller, H., Bull, J., 2011. The performance of farm animal assessment. UK Animal Welfare 20, 69–78.

Sant'Anna, A.C., Paranhos da Costa, M.J.R., 2011. The relationship between dairy cow hygiene and somatic cell count in milk. Journal of Dairy Science 94, 3835–3844. https://doi.org/10.3168/jds.2010-3951.

Schlageter-Tello, A., Bokkers, E.A.M., Groot Koerkamp, P.W.G., Van Hertem, T., Viazzi, S., Romanini, C.E.B., Halachmi, I., Bahr, C., Berckmans, D., Lokhorst, K., 2015. Relation between observed locomotion traits and locomotion score in dairy cows. Journal of Dairy Science 98, 8623–8633. https://doi.org/10.3168/jds.2014-9059.

Smith, B.P., 2014. Large Animal Internal Medicine. Mosby Elsevier, St Louis, MO, USA.

Stangaferro, M.L., Wijma, R., Caixeta, L.S., Al-Abri, M.A., Giordano, J.O., 2016. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part I. Metabolic and digestive disorders. Journal of Dairy Science 99, 7395–7410. https://doi.org/10.3168/jds.2016-10907.

Stygar, A.H., Frondelius, L., Berteselli, G.V., Gómez, Y., Canali, E., Niemi, J.K., Llonch, P., Pastell, M., 2023. Measuring dairy cow welfare with real-time sensor-based data and farm records: a concept study. agriRxiv 2023, 20230235216. https://doi.org/10.31220/agriRxiv.2023.00185. Published online by agriRxiv 2 June 2023.

Stygar, A.H., Gómez, Y., Berteselli, G.V., Dalla Costa, E., Canali, E., Niemi, J.K., Llonch, P., Pastell, M., 2021. A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. Frontiers in Veterinary Science 8, 177. https://doi.org/10.3389/fvets.2021.634338.

Stygar, A.H., Krampe, C., Llonch, P., Niemi, J.K., 2022. How far are we from data-driven and animal-based welfare assessment? A critical analysis of European quality schemes. Frontiers in Animal Science 3, 874260. https://doi.org/10.3389/fanim.2022.874260.

Teng, K.-T.-Y., Devleesschauwer, B., Maertens De Noordhout, C., Bennett, P., McGreevy, P.D., Chiu, P.-Y., Toribio, J.-A.L.M.L., Dhand, N.K., 2018. Welfare-Adjusted Life Years (WALY): A novel metric of animal welfare that combines the impacts of impaired welfare and abbreviated lifespan. PLoS One 13, e0202580.

Tucker, C.B., Jensen, M.B., De Passillé, A.M., Hänninen, L., Rushen, J., 2021. Invited review: Lying time and the welfare of dairy cows. Journal of Dairy Science 104, 20–46. https://doi.org/10.3168/jds.2019-18074.

Tuyttens, F.A.M., de Graaf, S., Andreasen, S.N., de Boyer des Roches, A., van Eerdenburg, F.J.C.M., Haskell, M.J., Kirchner, M.K., Mounier, Luc, Kjosevski, M., Bijttebier, J., Lauwers, L., Verbeke, W., Ampe, B., 2021. Using expert elicitation to abridge the Welfare Quality® protocol for monitoring the most adverse dairy cattle welfare impairments. Frontiers in Veterinary Science 8, 634470. https://doi.org/10.3389/fvets.2021.634470.

Tuyttens, F.A.M., Molento, C.F.M., Benaissa, S., 2022. Twelve Threats of Precision Livestock Farming (PLF) for Animal Welfare. Frontiers in Veterinary Science 9, 889623. https://doi.org/10.3389/fvets.2022.889623.

Utriainen, M., Pastell, M., Rinne, M., Kajava, S., Myllymäki, H., 2019. Sensor Technologies in Dairy Farms in Finland. In: O'Brien, B., Hennessy, D., Shalloo, L. (Eds.), Precision Livestock Farming 2019. ECPLF, Teagasc, Ireland, pp. 98–104.

Visual Crossing Corporation, 2021. Visual Crossing Weather. Retrieved on 10 June 2022 from www.visualcrossing.com.

Vokey, F.J., Guard, C.L., Erb, H.N., Galton, D.M., 2001. Effects of alley and stall surfaces on indices of claw and leg health in dairy cattle housed in a free-stall barn. Journal of Dairy Science 84, 2686–2699. https://doi.org/10.3168/jds.S0022-0302(01)74723-6.

Welfare Quality®, 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, Netherlands.